

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
HAYSTACK OBSERVATORY
WESTFORD, MASSACHUSETTS 01886

Telephone: 978-692-4764
Fax: 781-981-0590

March 29, 2004

TO: Distribution
FROM : David Lapsley
SUBJECT: 22 March 2004 e-VLBI telecon summary

Attendees:

Bill Fink, Lee Foster, Pat Gary, Paul Lang, Mary Shugrue – GSFC
Dennis Baron – MIT
Jerry Sobieski – University of Maryland Mid-Atlantic Cross-roads
Charles Yun – Internet2
Terry Gibbons – Lincoln Laboratory
Kevin Dudevoir, Hans Hinteregger, David Lapsley, Arthur Neil, Alan Whitney Haystack

This telecon is one of an ongoing series of telecons to prepare for 10 gigabit/sec e-VLBI demonstrations between NASA GSFC and MIT Haystack Observatory using a combination of network facilities including all or part of GlowNet, Bossnet, ISI-E, SuperNet, Max and GSFC/HECN.

ACTION ITEMS ARE HIGHLIGHTED IN RED.

Bossnet

Alan Whitney: Jerry Sobieski here at Haystack, spent time yesterday talking about how to make OC48 connection to Bossnet at Lincoln and Washington end. Detailed discussion with David, Terry, Jerry and Tom about a week ago. Looks like things are coming together, but still a few details to work out.

Jerry Sobieski: doing a sanity check on the WaveShifters David has spec-ed out. Need to get confirmation on ATDnet guys that the wavelength is acceptable to use end to end. I've sent Linden that suggestion, not heard back from him yet. Once this is ironed out, can give David the wavelength and order the Wavelength shifters. A few questions on network diagram. May need to find out more about Cisco 1500 and MUX/DEMUX at Lincoln. Terry, is there a Cisco 1500 at Lincoln.

Terry Gibbons: this equipment is on GlowNet. Russ Roberge knows more about this equipment. I'll discuss with him at 3:00 pm meeting.

Jerry: there is a mux that front ends the run down to Washington.

Terry: in theory should be a first wave switch that they will send up to us. In practice, using home-built optical equipment.

ACTION: Terry and Jerry to discuss offline.

Jerry: still working on getting all fiber in place. Contracting issues taking time, not technical. Contracts should come to closure in next week or 10 days. Fiber probably up in 30 days. Hoping that all objects and fibers installed before the end of May.

Alan: we are expecting a Juniper M10 router next week. Will be shipped on the 25th of March. Preparing to order Waveshifters. Want to make sure that we are ordering all the right equipment, and that all of the specifications are appropriate.

Pat Gary: No changes at GSFC relative to existing network diagrams. Small level tests/constant polling across GGAO link every hour or so.

Alan: Jerry, would you like to talk about DRAGON?

Jerry: 30 second summary of DRAGON. Worked with Alan and David last spring to put together a proposal to use the e-VLBI project as an application that would demonstrate dynamic optical resource allocation. DRAGON is an acronym for Dynamic Resource Allocation over GMPLS Optical Networks. Collaboration between MAX and the University of Maryland, ISI in Arlington Virginia, George Mason University, MIT Haystack and Movaz Networks (optical network equipment provider out of Atlanta). Four key areas being worked on:

1. Defining inter-domain requirements for progressing optical circuit requirements from one domain to the next. Issues that are not well understood and not standardized. Tom Lehman is heading up this area.
2. Virtual Label Switched Router: takes a group of non-GMPLS routers and hides them under a Virtual LSR. This concatenates islands of GMPLS activity transparently. Use GMPLS, signaling and routing standards for doing wavelength and Ethernet and SONET type circuits. Can extend this out beyond a given network. Important aspect. A lot of technical difficulties in this area.
3. Application Specific Topologies Definition Language: allows the creation and definition of Application Specific topologies. In a way that they can be re-created at different times. Schedule them in advance, re-allocate the terminal nodes at different locations. Can address the issue that Matt Mathis talks about – the “Wizard Gap” – gap between what the network guys know the network can do and what the application guys know what the network can do. Not enough overlap between the two to take advantage of either. Trying to formalize the network performance from an application stand point. And then be able to instantiate the topology for something like e-VLBI or a campus LAN. Have some prototypes that are almost ready to start showing off. Simple concept but lots of detail.
4. Optical routing: partnership with Movaz. Have optical switches and OADMs that use GMPLS

to setup circuits within that environment. To the degree that these GMPLS stacks interoperate with other vendors' stacks, in theory should be able to connect networks together and setup circuits across a given domain. Interesting aspects: Movaz has given us pre-production equipment, the optical switching technology that they are developing in their labs. We have all optical switches at UMD (installed in December). That allows us to switch wavelengths all optically. This will grow over the next three years. Already in discussions about what sort of technology we want to see (tuneable lasers, optical alien wavelengths channel monitors). Bijan Jabbari (faculty at GMU working on optical routing algorithms). Lots of low level issues (e.g. dispersion, attenuation, blocking etc.) need to integrate them into the routing protocol.

A couple of key applications, one of which is e-VLBI. Sometime in the next 3–6 months want to have a demonstration of this with e-VLBI. Also have UltraGrid work at ISI: taking AccessGrid and extending it into HDTV. Doing that over gigabit ethernet. Want to be able to establish video connections using circuits rather than just packet streams. Questions?

Kevin Dudevoir: How does ASTDL in DRAGON mesh with Dave Lapsley's EGAE. Which allows the user to specify a networking profile.

David Lapsley: I can talk briefly about EGAE and then you might like to talk about DRAGON. EGAE is the Experiment Guided Adaptive Endpoint that we are working on here. It is basically an intelligent end system that will act as an interface between VLBI/Mark5 type end equipment and an IP network. One of the characteristics of the EGAE is that you can specify the behavior of the EGAE in a high level profile that will allow you to specify things like tolerable cell loss, tolerable delay, bounding requirements on the length of the transfer, things like that which can be specified by scientists at a high level. So that the EGAE can transfer the data in a some sense optimal fashion between the Antenna and the correlator. The way I would see it, EGAE is more like a way to optimize end to end transfer of data. Whereas ASTDL, if my understanding is correct, is more about defining the topology, setting up the connections in the network, quality of service.

Jerry: hesitate to open up the QoS per-se, but can't avoid it. ASTDL is an attempt to formalize what the application looks like in terms of the network requirements. Appears that we will be having the ability to specify some sort of bounding on the performance and use that to define the types of framing and types of circuits that need to be established. The way I envision ASTDL. It needs to be very simplistic in its first approximation so that you don't overwhelm people that don't know much about it. That simplistic function is defaulting a lot of specifications, if you don't specify a framing type, it assumes a certain framing type, as for latency and jitter requirements. Part of the value for ASTDL is being able to make sure that this is deterministic. So if topology is run today, and tomorrow, you will get the same performance out of the network. I'm not familiar with EGAE, so perhaps we can talk about this over the next day or so.

e-VLBI Experiments

Alan: crossed an important milestone two or three weeks ago. Achieved full real-time processing of data from two antennas: Westford and GGAO. Done without any disc buffering. Only a few seconds

of electronic buffering. Through the efforts of a number of people down at Goddard and people here working on correlator software etc. It is a first (outside of Japan). The Japanese have done this with dedicated ATM circuits. It is the first real-time over IP I think. Europeans were very kind about it and said that maybe we could help them do the same thing. We had to do it at a pretty low data rate, because the only connection we have now is a 100 Mbps standard Internet connection to Haystack. We did this experiment at 32 Mbps. It wasn't impressive by data rate, but limited by the connection we had. We don't think there is any reason why it should not work at much higher data rates if we can get the data into the correlator. Later this week we are planning a similar experiment with Onsala Sweden. If that is successful, that will be the first international real-time e-VLBI. We have been having some difficulties with the link to Onsala. Nominally it is a 1 Gbps link, but we have been seeing data rates of 10 Mbps TCP. David has been working with people in Europe and the US along various points in the path and with the Internet2 people who have placed a machine in Europe to help with the isolation of the problem. We are struggling to get the data rate and understand where the bottleneck is and how to clear it. If we can get to 32 Mbps then we will be alright for doing this. Due to constraints of the hardware that is the minimum we can do.

Pat: Alan, you said that there was no disc capture at 32 Mbps. Can you slow the rate at the Antenna?

Alan: yes, we can slow it down to 32 Mbps as the minimum. Then it goes in octaves: 64 Mbps, 128 Mbps, etc...

Jerry: did you try doing it at 64 Mbps?

Kevin: averaged around 49 Mbps with single stream testing. First we had a very poor 5 Mbps and then chased this down to Haystack's connection. Haven't pursued it any further. Transfer data to Kame at ISI-E and then use an application level relay to transfer the data to Haystack.

Alan: it was a jury rigged experiment, but it did work and allowed us to claim the high ground.

Kevin: if you fall outside of the buffers at the correlator (we have 0.5 GB buffer) very difficult to recover.

Jerry: what is the best data rate you could manage?

Alan: we think we could manage 512 Mbps.

Kevin: we have done 256 Mbps auto-correlation from Westford.

Alan: we placed a Mark5 system at ISI-E: 1. so that we can transfer data there while Bossnet is down. 2. so that data can be collected there that USNO people can use "sneakernet" to come and pickup data on a daily basis. Still sorting this out, that would be from a station in Hawaii and a station in Germany. Using this quasi-real-time technique would allow us to get the turnaround time under a day (typically 3-4 days). Would be a big improvement on getting rapid turnaround. We are still working with the navy and pushing the navy to try and get a high speed connection to USNO. We are also hoping that

if the “sneakernet” proposal is successful that will generate enthusiasm at USNO at the higher levels to free up money to make the high speed connection to USNO itself. Those two things have happened here over the last couple of weeks.

Kevin: on the Kokee link to Hawaii. Clyde Cox (station chief) met with Roger Hall (ITT contractor) on Friday. They are getting more aggressive about trying to resolve the poor performance we are getting (<10 Mbps in the Haystack to Kokee connection over the OC3 link). They will be doing some dedicated testing this week. If they can't resolve this, they would be willing to host a test point at PMRF. We can get about 30 Mbps using Tsunami or UDT rate-based protocol. Supposed to be TCP friendly, but not sure. UDP we can get close to 100 Mbps over a 155 Mbps link with fairly low loss in the direction from Kokee to Haystack. In the other direction, we get high packet loss anywhere from 4-26% packet loss. Best we can do is 10 Mbps TCP.

David: what is the UDP loss from PMRF to Haystack?

Kevin: very low, 0.01%.

[Request to speak up]

David: I was just asking Kevin what the packet loss is from PMRF to Haystack. He said that it was 0.01%. Actually quite high for TCP, on a 120 ms round trip time that can cut down 10 Mbps. That is what we are seeing from Sweden.

Performance Testing and Monitoring

David: the main thing I have been working on is testing the performance from Onsala. One of the things that has been useful for doing this is the bwctl tool that Internet2 has released recently. That has been very useful for automating the tests. One of the big advantages is that now we have our encryption keys, we sent our encryption keys in to Internet2, we can now test to some of the performance servers that are located on the Internet2 backbone, as well as to the server that Alan had mentioned which is in the UK on the GEANT network. I don't know if anybody has had a chance to check it out, but if you do get a chance, have a look at the Internet2 website. They have a nice page up that describes the architecture and you can download the software and try it out. I think it is definitely worthwhile:

<http://e2epi.internet2.edu/bwctl/>

At the moment, that is the main thing that I have been working on. It's basically a couple of daemons that act as daemons around iperf. So you run the daemons on your end system and then by starting up a client, you can initiate an iperf test that is subject to authorization and resource limitations. You can also parameterize the tests. I'm not sure what their plans are to expand it, but if you get a chance, Bill, you might like to see if nuttcp could fit into that. Nuttcp has a lot of that functionality in it, and it nice to test with, but I guess at the moment they are using Iperf for doing that. So that might be something to look at.

Alan: Charles do you have anything to add to this?

Charles Yun: No, I think that pretty much covers it from the bwctl side.

Alan: David has been quite impressed with bwctl. He has been using it quite a bit.

Pat: On performance, NC state has developed a new transport protocol: BIC TCP, Binary Increase Congestion TCP. According to the article, it sounds as though it may be superior to the others. I was wondering if David might want to look at this.

David: Thanks Pat, I'll take a look at it. I had heard about this, but don't know much about the details. I'll check it out.

Pat: I will email out to the list a reference.

Alan: Pat, did Bill Wildes contact you about expanding the link to GGAO?

Pat: not really, one brief telephone contact where I was trying to make him aware that in a month or two, we might be suggesting certain kinds of upgrades. Give him a heads up and trying to get a sense of whether there was any kind of budget in this fiscal year to respond to that sort of change. Things open right now. Not sure exactly what the status is. Will continue to explore these options and when we have something that makes sense, run it by him, and he could then take that and see if he can get support for it.

Alan: I spoke with Bill about this a couple of weeks ago. He said he currently doesn't have any budget. I suggested he might speak with you about this, but I am not sure where it went from there. I will mention it to him.

Pat: We own a Force10 E300 with a pair of 10 GigE blades and 2 x 12 port GigE's (will be replaced with a 48 port card). Force10 loaned us a mirror image of this system. We put them back to back with 2 x 10 GigE blades. Bill Fink and Paul Lang worked on a scheme using two pairs of G4 Macs. Between one pair they generated a 1 GigE stream into one of the 1 GigE ports, but then looped it via VLANs so that they would traverse the 10 GigE blades to go to the other one and then patch the switch back into itself. Through the cross-strapping with another pair of G4 macs pushing in the opposite direction. Through the pair of 10 GigE cards, able to generate 40 Gbps of bandwidth moving between the two switches. The fun part was running it for a weekend (72 hours), for the duration of this test they were able to ship over 1 petabyte of data.

Very pleased with the switching holding up, all ports working at line speed. Able to bring out the capabilities of VLANs, link aggregation and get some sense of latency. Looking forward to moving one switch to the MAX and test across the existing CWDM links. In Goddard, have taken one of the computer clusters (Thunderhead – listed as 106th ranked supercomputer). Temporarily split it into two logical machines and wired up 16 heads with 1 GigE interfaces that will go to one of these switches and transmit across a 10 GigE backplane to another 16 nodes or heads in another building. Eventually to

a cluster across the country. We can send you some of this data, Paul Lang worked over the weekend, Bill had worked on some charts, Paul Lang had put some further intro material with that and we have a show and tell we will do with the Force10 people tomorrow. They asked us to be speakers at a seminar they are hosting in the Washington, mid-Atlantic region (informational seminar). After that, we will email the data to you.

Alan: what was the average rate for the 10 Gbps test?

Pat: from the MRTG monitoring, peak is 100% utilization on the 10 G ports: 9.9991 Gbps. Average is 9.8 Gbps.

Pat: I want to talk about the National Lambda Rail. Not to take anything away from the continuing efforts to have a Bossnet link, Abilene, Internet2 etc. NLR proposes to have itself extended to Boston. I'm not sure about the time frame. Shows itself as being deployed through New York and Boston. If it's going to Boston, I would imagine it is going to someplace like MIT.

Alan: I'm not sure about this. Last we heard there were no plans about bringing NLR to Boston. We looked into doing that about a year ago and got a number of \$370K of NRE, plus a substantial monthly outlay.

Pat: in this case, I am looking at a diagram on their home page that shows the first round of links. From DC focused on going south to Atlanta in the first phase (from now to September). Atlanta to DC, out to Carnegie Mellon, Chicago, Seattle area and then down the west coast to San Diego.

Alan: a few weeks ago, I ran into Ed Fantegrossi from IEAAF. He said that he had just got a donation from ATT from New York to Boston. Perhaps NLR has jumped on this.

Pat: NASA has some interest in NLR. Goddard wants to get to the west coast: University of California, San Diego, Scripps institute of Oceanography. Ames research colleagues, see value in having JPL, Ames and Goddard all connected to NLR as a way that NASA can participate. Larry Smarr invited a couple of NASA people to meet him at a meeting in Denver, with the CTO and several department heads from Level 3. He made an interesting pitch, asking them for additional lambdas (above the huge investment they have already made to NLR) which might be brought up prior to September. He's interested in lambda's between San Diego and Chicago. This would enable him to get to Chicago directly. He was willing in his pitch, to fold in the NASA interest, having Goddard connected through to Chicago/StarTAP or direct to San Diego. Myself, gentlemen from Ames and my boss were there confirming that this would be very helpful to NASA, not as customer, but as research partners/collaborators. Level 3 seems very receptive to having us there, indicated that they thought they could do something for us, looking at possibility of getting fiber, working with regional fiber folks, to possibly get connections into some of the NASA centers. Same lambdas and/or fibers could get rolled up into NLR or get phased out. Looks like some positive opportunities, would help me with a local air gap, that even the DRAGON project has with connecting to McClean. I'll tell you more as soon as level 3 puts it in writing.

Alan: that is very encouraging.

ACTION ITEM: Sometime in the next day, we will meet with Terry, Jerry and David to meet and iron out the last few details of the Bossnet connection.

ACTION ITEM: As soon as this is done, Alan will update the network diagram.

Next telecon is scheduled for Monday, 12th April 2004 at 2 pm EST.

cc: Steve Bernstein, LL
Jim Calvin, LL
Rick Larkin, LL
Lorraine Prior, LL
Peter Schulz, LL
Leslie Weiner, LL
Herbert Durbeck, GSFC
Bill Fink, GSFC
Lee Foster, GSFC
Pat Gary, GSFC
Andy Germain, GSFC
Chuck Kodak, GSFC
Kevin Kranacs, GSFC
Paul Lang, GSFC
Aruna Muppalla, GSFC
Mary Shugrue, GSFC/ADNET
Bill Wildes, GSFC
Dan Magorian, UMCP
Tom Lehman, ISI-E
Jerry Sobieski, MAX
Guy Almes, Internet2
Charles Yun, Internet2
Richard Crowley, Haystack
Kevin Dudevoir, Haystack
Hans Hinteregger, Haystack
David Lapsley, Haystack
Arthur Niell, Haystack
Joe Salah, Haystack